On the Effect of Technology Scaling on Variation-Resilient Sub-Threshold Circuits

Nele Reynders[a,*], Wim Dehaene[a]

[a]*Dept. of Electrical Engineering (ESAT-MICAS), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium*

Abstract

This paper studies the impact that CMOS technology scaling has on circuits operating in the ultra-low-voltage region. Sub-threshold circuits are an attractive option for energy-constrained applications, but the influence of scaling on the energy consumption has not been studied thoroughly on on-chip ultra-low-voltage implementations. This paper aims to provide an answer to the benefits and disadvantages of scaling on such implementations. First, an equation to determine the minimum feasible supply voltage for digital circuits is derived. Out of this equation, a theoretical minimum as well as a practical minimum supply for a specific technology can be calculated. Second, a 16-bit Multiply-Accumulate Unit is selected as a test vehicle to study scaling effects. This test vehicle is designed, processed and fully measured in both a 90 nm and a 40 nm CMOS technology. An extensive comparison between the measurement results of both designs allows to clearly examine the different technology scaling trade-offs.

1. Introduction

CMOS scaling has been driven by the increased performance that is obtained for digital systems. However, many applications do not require such a high circuit speed. Instead, the energy consumption becomes the critical parameter. A wide range of these applications exists, e.g. sensor networks, RFID tags and biomedical signal processors. Ultra-low-voltage circuits can be a solution for such energy-constrained applications that are less stringent on speed requirements [1]. Lowering the supply voltage V_{dd} under or near the threshold voltage V_{T} enables large reductions in energy consumption, at the disadvantage of a simultaneous increase in circuit delay. Although few implementations only require operating frequencies in the kHz-range, to allow a more widespread use of ultra-low-voltage circuits, higher operating frequencies well within in the MHz-range are required.

The impact of CMOS technology scaling for digital circuits operating at the nominal supply has been extensively studied, through simulations, on-chip implementations and measurements. The influence of scaling on circuits operating in the weak inversion region has received some attention, but until now, this attention has been limited to device-level studies and circuit-level simulations.

*Corresponding author

Email addresses: nele.reynders@esat.kuleuven.be (Nele Reynders), wim.dehaene@esat.kuleuven.be (Wim Dehaene)

Previous work investigated the effect of scaling on device-level and proposed different scaling strategies. [2] performed a model study of sub-threshold transistors to investigate the implications of device scaling on sub-threshold operation from 90 nm down to 32 nm technology nodes. An alternative scaling strategy was proposed to help sub-threshold circuits to reliably scale to nanometer technologies. In [3], devices were redesigned specifically for sub-threshold operation. An optimized transistor structure to improve sub-threshold the circuit delay and the power delay product was proposed. The impact of technology scaling for nodes from 90 nm to 22 nm was examined in [4] and strategies for increasing the robustness of sub-threshold circuits were proposed.

Some prior works also performed simulations to examine the impact of technology scaling on ultra-low-voltage logic circuits, although this mostly consisted of simple circuits. In [5], simulations of a ring oscillator were used to validate an analytical approach for studying the effect of technology scaling and variability on performance of ultra-low-power integrated systems. The effects of process variations were exhaustively examined to study the sensitivity of a circuit in presence of these variations. [6] investigated the impact of technology scaling on sub-threshold logic in nodes from 0.25 µm to 32 nm CMOS. A circuit-level simulation of a benchmark 8-bit multiplier was used to study the scaling effects, first using predictive technology models (PTM) and then validated by industrial models.

To the author's knowledge, only [7] presented measured results: two test chips were fabricated in a 130 nm and a 65 nm technology, consisting of a 1000-stage inverter chain and a 41-stage ring oscillator, both operating at ultra-low voltages. The measurements of these two simple circuits were used to validate a body biasing technique to adaptively balance the pMOS and nMOS transistors in strength. No papers have been published that present the design and measurements of a full digital system, implemented in different CMOS technology nodes. Moreover, a significant amount of the previous work did not use industrial models, but rather relied on PTMs (e.g. [2, 4, 5]). PTMs are reasonably accurate transistor models that benchmark future generations of technologies and are therefore a useful resource for early circuit design research [8]. Although PTM simulations are very suitable to investigate scaling trends, they do not provide the same value as designing with industrial models, followed by manufacturing and measuring the designed chip.

This paper aims to fill the hiatus between simulations and measured, confirmed results. An extensive digital circuit has been designed, processed and measured in both a 90 nm [9] and a 40 nm CMOS technology. The test vehicle that was used to study the effect of technology scaling on ultra-low-voltage circuits, is a 16-bit Multiply-Accumulate Unit (MAC). The MAC was chosen since it is a block that is very frequently used in DSP designs and because it is a complex block that includes feedback. Since the MAC is the critical component of DSP designs, it is also possible to design a processor that achieves similar ultra-low-voltage characteristics as this MAC with the same design principles [10]. Important to note is that the aim was to design variation-resilient circuits which are able to operate at both very low energy consumptions and $n \times 10$ MHz-speed to increase the industrial relevance of ultra-low-voltage circuits. To conclude, this

paper focuses on providing a scaling analysis which is based on the design and measurements of a large ultra-low-voltage circuit.

Section 2 gives an overview of previous theoretical studies to find the fundamental limit for the lowest feasible supply voltage $V_{\mathrm{dd},min}$ and provides a new, practical equation to determine $V_{\mathrm{dd},min}$ for a specific technology. Section 3 gives a comparison between the used technologies, while Section 4 covers the design of the 16-bit MAC. Section 5 explains in detail the transistor-level implementation of the different logic components. The design changes that were necessary for the technology scaling are also addressed. In Section 6, the implementation of the timing is described, with a specific focus on the different design decisions necessary for both technologies. Section 7 presents the results obtained from the measurements of both chips.

2. Theory of Sub-threshold Operation

Before explaining the design and implementation of the MAC test case in the ultra-low-voltage region, this section first provides a practical expression that estimates the minimum possible supply voltage that can be expected for a certain technology. Previous research has focused on theoretically finding the fundamental limit for the lowest operating voltage for CMOS technologies. Already in 1972, [11] studied the minimum usable supply of an inverter, with the requirement that the inverter should have sufficient maximum gain at $V_{\mathrm{dd}}/2$ to be usable in a digital circuit. Based on measurements in a technology available at that time, the authors estimated that the minimum usable V_{dd} would have a value of about $8kT/q$ (where k is the Boltzmann constant, T the absolute temperature and q the electrical charge of an electron), or 207 mV at 300 K. In 2001, [12] proposed another theoretical limit of the lowest operable supply. To achieve this V_{dd}, the nMOS and pMOS off-currents must be equalized. Following this requirement, the ideal supply limit of $4kT/q$ is proposed, which is 103 mV at 300 K.

These are all theoretical limits that predict the lowest possible supply voltage of CMOS digital circuits. However, they are not practical limits that take into account the specific details of the technology at hand. Therefore, we have derived a practical limit for the minimum feasible supply from the equations listed below. The basic equation for the current flow in the weak inversion region consists of an exponential relationship with V_{GS}:

$$I_{\mathrm{DS}} = I_0 \cdot \exp\left(\frac{V_{\mathrm{GS}} - V_{\mathrm{T}}}{n \cdot V_{\mathrm{th}}}\right)\left(1 - \exp\left(\frac{-V_{\mathrm{DS}}}{V_{\mathrm{th}}}\right)\right) \tag{1}$$

where V_{th} is the thermal voltage ($V_{\mathrm{th}} = kT/q = 26$ mV at room temperature 300 K), n is a process-dependent parameter and I_0 is the current when $V_{\mathrm{GS}} = V_{\mathrm{T}}$. I_0 is dependent on process and device geometry [13, 14]. The third term incorporates the current roll-off, and only has an influence when V_{DS} drops to within a few multiples of V_{th}. Taking the ratio of the on-current I_{on} at $V_{\mathrm{GS}} = V_{\mathrm{dd}}$ and the off-current I_{off} at $V_{\mathrm{GS}} = 0$

gives:

$$\frac{I_{\text{on}}}{I_{\text{off}}} = \frac{I_0 \cdot \exp\left(\frac{V_{\text{dd}}-V_{\text{T}}}{n \cdot V_{\text{th}}}\right)}{I_0 \cdot \exp\left(\frac{-V_{\text{T}}}{n \cdot V_{\text{th}}}\right)} = \exp\left(\frac{V_{\text{dd}}}{n \cdot V_{\text{th}}}\right) \tag{2}$$

In both I_{on} and I_{off}, $V_{\text{DS}} = V_{\text{dd}}$ and therefore V_{DS} dependencies can be omitted. In (2), a direct relationship between the variables V_{dd}, I_{on} and I_{off} is obtained since V_{th} is fixed for a certain temperature and n is fixed for a certain technology. An equation for the supply voltage can be derived:

$$V_{\text{dd}} = \ln\left(\frac{I_{\text{on}}}{I_{\text{off}}}\right) \cdot n \cdot V_{\text{th}} \tag{3}$$

The value of n is affected by depletion region characteristics [15] and is equal to 1 for an ideal transistor, but unfortunately larger than 1 for actual devices. It is typically in the range of 1.3-1.7 for CMOS processes [14]. Since it is difficult to accurately determine n for a certain technology, the link with the so-called sub-threshold slope S_S will be made. The sub-threshold slope is defined by the amount by which V_{GS} must be increased in order for the weak inversion current I_{DS} to be increased by one order of magnitude. It is expressed in mV/decade [15]:

$$S_S = n \cdot V_{\text{th}} \cdot \ln(10) \tag{4}$$

Substituting n in (3) by using (4), results in:

$$V_{\text{dd}} = \frac{\ln\left(\frac{I_{\text{on}}}{I_{\text{off}}}\right) \cdot S_S \cdot V_{\text{th}}}{V_{\text{th}} \cdot \ln(10)} = \log_{10}\left(\frac{I_{\text{on}}}{I_{\text{off}}}\right) \cdot S_S \tag{5}$$

This result shows that for a certain CMOS technology (and thus for a certain S_S), the minimum supply V_{dd} is only dependent on the minimum $I_{\text{on}}/I_{\text{off}}$ current ratio. This equation makes it possible to derive a practical as well as a theoretical limit for the minimum feasible supply voltage for a circuit operating in the weak inversion region. The V_{dd} dependence of the $I_{\text{on}}/I_{\text{off}}$ ratio is logical: the lower V_{dd}, the lower I_{on} will be obtained, and the lower the current ratio will become. From experience, a fair minimum value for the $I_{\text{on}}/I_{\text{off}}$ current ratio is 50. A lower value of the current ratio becomes problematic, since the circuit robustness in the presence of variations will be compromised. A theoretical limit for the minimum supply voltage can be found through the theoretical lower bound of the sub-threshold slope S_S. In the ideal case, n is equal to 1 and therefore the minimum S_S is equal to 60 mV/decade at room temperature. The theoretical $V_{\text{dd},min}$ can then be calculated to be 101 mV.

However, although devices with an ideal sub-threshold slope are optimal for sub-threshold applications [16], typical S_S values for a bulk CMOS process range from 70 to 120 mV/decade [17], well above the theoretical lower bound. Unfortunately, technology scaling has a bad impact on S_S because it is proportional to the gate-oxide thickness T_{ox} which does not scale in proportion to the physical gate length. Scaling of T_{ox} actually slows down starting from the 130 nm node to limit gate leakage [6]. A comparison of industrial publications in [8] indicated that T_{ox} has been reduced by about 10 % per generation between the 130 nm

CMOS Technology	90 nm	40 nm
S_S [mV/decade]		
nMOS	93.0	98.1
pMOS	86.1	109.9
$V_{\mathrm{dd},min}$ [mV]		
nMOS	**158**	166
pMOS	146	**186**

Table 1: Measured sub-threshold slope and resulting $V_{\mathrm{dd},min}$.

and the 40 nm technology nodes [2]. For the technologies at hand, T_{ox} decreases with about 13 % between the 90 nm and 40 nm nodes. As a result, S_S degrades as function of technology scaling.

Measurements of transistors placed next to the MAC in both technologies confirm this trend (see Table 1): S_S degrades 5.5 % going from the 90 nm to the 40 nm technology for an nMOS transistor and 27.6 % for a pMOS. The practical limit of $V_{\mathrm{dd},min}$ can then be calculated as the maximum $V_{\mathrm{dd},min}$ per technology, resulting in 158 mV for the 90 nm technology and 186 mV for the 40 nm technology. To conclude, by using a practical limit for the $I_{\mathrm{on}}/I_{\mathrm{off}}$ current ratio (thus taking into account both on-current and leakage current) and a value of S_S (which can be obtained through simulations or measurements), we have obtained a straightforward manner to calculate the practical minimum feasible supply voltage for digital circuits operating in the weak inversion region in a certain CMOS technology.

3. Technology Comparison

Before thoroughly explaining the detailed design in both technologies, it is important for the MAC design to analyze their similarities and differences. To be able to make a fair and consistent comparison, both CMOS technologies consist of similar performance equivalents. In order to obtain a higher current at low supply, not the low leakage equivalent of the technology nodes is chosen, but the general purpose or standard performance equivalent is used in both cases. For the same reason, all transistors use a similar threshold voltage option, i.e. low-V_{T} (LVT) transistors. With technology scaling from 90 nm to 40 nm CMOS, both the nominal supply voltage V_{dd} and the transistor threshold voltage V_{T} decrease slightly. Since V_{T} decreases and the $W_{\mathrm{min}}/L_{\mathrm{min}}$ ratio does not remain constant but instead increases with a factor of 2 (for the technologies at hand), one would expect the transistor current I_{DS} to increase for a certain supply. Therefore, the delay t_d should reduce, causing circuits to function at a higher speed. The dynamic energy consumption is expected to decrease with the third power of the scaling factor. The total chip area will decrease, as well as the cost. These are all advantages of scaling. However, there are also downsides to scaling. First, the leakage current will increase exponentially due to the reduced V_{T}. Second, because of the reduced transistor dimensions, transistor variability will have a higher impact and will thus become much more important.

(a) (b)

Figure 1: Technology model comparison: (a) On-current I_{on} for a minimal sized nMOS and (b) Propagation delay t_p for a regularly sized inverter ($W_{\text{nMOS}} = W_{\text{min}}$ and $W_{\text{pMOS}} = 3.W_{\text{min}}$) as function of V_{dd}.

When looking at technology scaling from an ultra-low-voltage perspective, all of the advantages remain. In fact, because of the decreasing V_{T}, circuits should become much faster for the same extremely low supply voltage. Since speed performance is often an issue in sub- and near-threshold designs, technology scaling becomes an even more attractive option for such designs. However, the disadvantages of scaling have an even higher effect in ultra-low-voltage designs. Due to the increased leakage and the aforementioned degradation of the sub-threshold slope, the $I_{\text{on}}/I_{\text{off}}$ ratios in the sub-threshold domain reduce to dramatically low values, thereby compromising circuit robustness. Moreover, the exponential sensitivity to variations combined with the overall increased variability results in problematic gate robustness. To conclude, technology scaling can definitely bring added value for ultra-low-voltage circuits, but it is imperative to take into account the low current ratios and the high variability at the time of design.

To obtain an understanding of the increased variations, a look at the impact of technology on V_{T} variation will be given. The threshold voltage is determined by the number and location of dopant atoms implanted in the channel region. Since the number of dopants is small in nanometer processes, the variation of V_{T} due to random dopant fluctuations becomes large [14]. The Pelgrom coefficient $A_{V_{\text{T}}}$ [18] provides a measure of the amount of variations. The Pelgrom coefficient for both technologies has been extracted out of 1000 Monte Carlo transistor simulations. For nMOS transistors, $A_{V_{\text{T}}}$ increases with 5.6 % when going from 90 nm to 40 nm, while the increase for pMOS transistors is as large as 41.4 %. These numbers clearly show the increased variations for advanced nanometer technologies.

An additional problem in the weak inversion region, is that the calibration of transistor models is not as reliable as it is in the strong inversion, nominal region. To give an example, Fig. 1(a) shows the simulation results of the on-current I_{on} of a minimal nMOS as function of V_{dd}. According to the models of the 90 nm and 40 nm CMOS technologies at hand, $I_{\text{on},40nm}$ is more than 2 times smaller than $I_{\text{on},90nm}$ for a supply of 200 mV, which directly contradicts the scaling effects that were expected. A similar unexpected behaviour can be seen in Fig. 1(b), which shows the propagation delay t_p of a regularly sized inverter for super-threshold operation. For the same 200 mV supply, $t_{p,40nm}$ is 56 % higher than $t_{p,90nm}$, resulting in an inverter which is more than 1.5 times slower. Fig. 1 also shows that in the nominal supply domain on the contrary, the

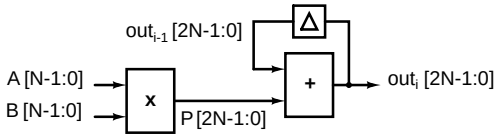


Figure 2: Standard MAC block diagram.

technology scaling expectations do apply. To conclude, these simulation results show that blindly relying on transistor models in the ultra-low-voltage domain is not recommendable.

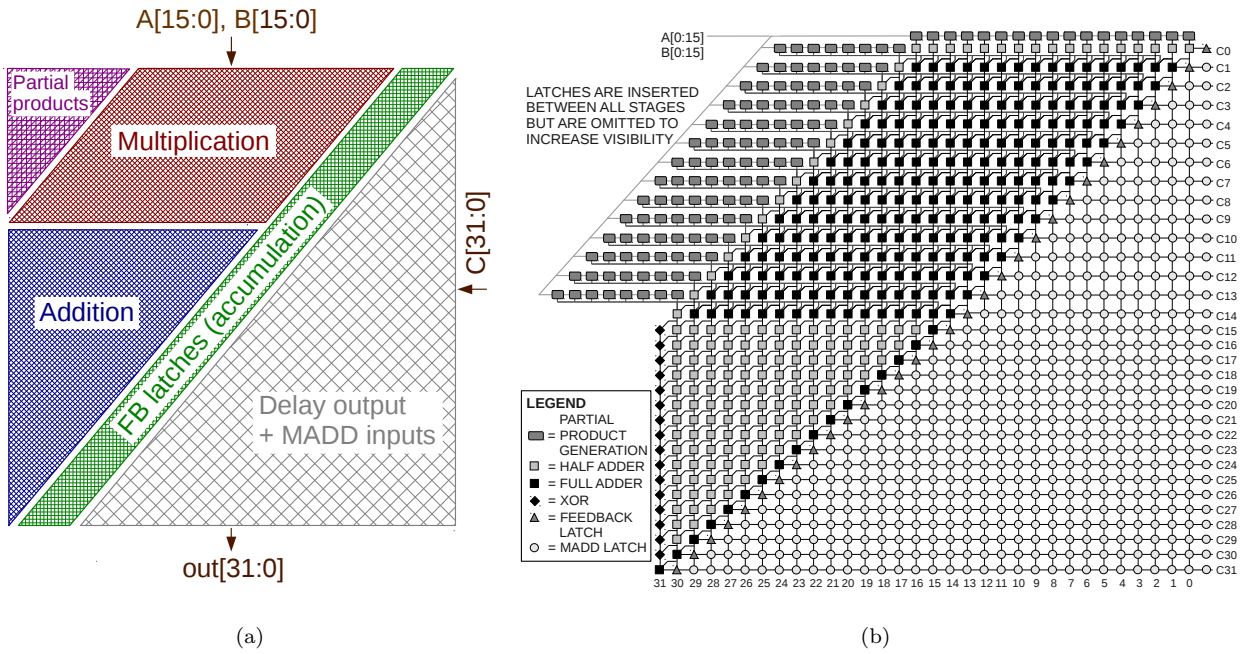
Some nuances should be given to the discussion of the reliability of the transistor models. Although *absolute* numbers are not very trustworthy (such as the exact energy consumption or propagation delay of a circuit), *relative* comparisons in the same technology using the simulation results are definitely valuable. Out of experience from previous designs (e.g. [19]), we also know that simulations to check functionality produce reliable results. Moreover, as can be seen from the variation analysis above, the intra-die simulations do show the expected increase in variability when going to a smaller technology and therefore such simulations do provide realistic results.

4. Design of the MAC unit

Now that the differences between both technologies are explained, this section will cover the design of the test case that was studied. The implemented chip cannot only perform Multiply-Accumulate operation, but the system is expanded so that it can also operate in two different modes: multiplier (MULT) and multiply-add (MADD) mode. The equation that represents the operation of the system is $out = A.B + C$ where A and B are N-bit binary numbers and C is a $2N$-bit number. The output out must be at least a $2N$-bit number. In MAC operation, the input C is in fact the previous output out_{prev}. In MULT operation, C is changed to zero, while C represents a third, $2N$-bit input for MADD mode.

From an architectural point of view, the design uses a pipelined architecture to increase throughput in order to achieve a high clock frequency. More specifically, a latch-based pipeline is utilized since latches tolerate time borrowing. Time borrowing is a very beneficial concept for sub-threshold designs due to the highly variable gate delays.

Standard MAC designs consist of a multiplier followed by an adder to perform the accumulation (Fig. 2). However, deep pipelining is not possible with such a standard design, because the maximal pipeline depth is only 2. To allow more pipelining, the structure of the MAC implementation has to be changed. Since the basic form of multiplication can be reduced to the addition of partial products, extending this addition step replaces the need for the separate accumulation step. This design thus consists of a multiplier which is extended with an interwoven accumulation structure. Fig. 3(a) shows a functional diagram of the implemented MAC where the interwoven diagonal accumulation is clearly visible. This accumulation is obtained



(a) (b)

Figure 3: Architecture of the 16-bit Multiply-Accumulate unit: (a) functional block diagram (b) detailed gate-level architecture.

through feedback (FB) latches which perform bit-by-bit feedback of the previous output. Not only does such an interwoven implementation allow to efficiently pipeline the architecture, it also significantly reduces the total delay for the multiply-accumulate operation to a delay slightly higher than needed for multiplication only. Another advantage is the much higher throughput that can be achieved through deep pipelining.

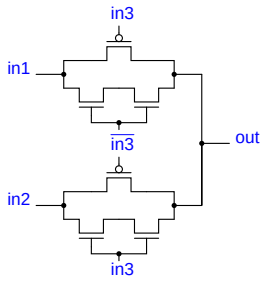
The implemented multiplier is based on the Modified Baugh-Wooley multiplier algorithm [20]. Fig. 3(b) gives the detailed gate-level architecture of the MAC. The MAC is able to work with two's complement numbers. The implemented MAC operation consists of 16-bit multiplication with 32-bit accumulation. On system level, the choice of operation mode is performed through two configuration bits that direct three timing control signals. Those timing signals are fed to the feedback latches which therefore enable the operation in the different modes. Calculations in every mode take the same amount of latency and there is no delay penalty for a multiply-add or multiply-accumulate operation with respect to multiplication.

5. Implementation of the MAC unit

This section handles the transistor-level implementation of the MAC. The differences in implementation between the 90 nm and 40 nm CMOS technologies are also addressed. Note that in both technologies, all transistors used in the chips are LVT devices, to be able to make a fair comparison.

5.1. Logic Gates

Careful design of logic gates is crucial if they should be able to efficiently work in the ultra-low-voltage region. Their topology not only has a large impact on the variation-resilience of the total design, but also



Logic gate	in1	in2	in3
OR	A	V_{dd}	B
NOR	$\overline{A}$	V_{ss}	B
AND	A	V_{ss}	$\overline{B}$
NAND	$\overline{A}$	V_{dd}	$\overline{B}$
XOR	A	$\overline{A}$	B
XNOR	$\overline{A}$	A	B
MUX	A	B	Sel
MUX bar	$\overline{A}$	$\overline{B}$	Sel

Figure 4: Logic gate design with generic Transmission Gate logic.

Technology	90 nm CMOS	40 nm CMOS
TG logic gate		
Sizing nMOS	2	1
Sizing pMOS	2	2
Inverter		
Sizing nMOS	1.5	2
Sizing pMOS	9	10

Table 2: Implementation details per technology.

on the delay, leakage power and active energy consumption. It has been shown that this trade-off is most optimal when using Transmission Gate (TG) logic, extended with transistor stacking [21]. TG logic is preferred because of its higher variation-resilience and lower contribution to leakage than standard CMOS logic. An extensive comparison between both logic families in the 90 nm CMOS technology is available in [21]. Performing the same analysis for the 40 nm technology produced very similar results and favored TG logic again, especially seeing the higher variability in this smaller technology. Fig. 4 shows the implementation of TG logic extended with nMOS stacking. A single generic TG logic block can function as 8 different logic functions, including non-inverting gates. The design of this generic block has to be optimized only once for the specific technology at hand [9]. To improve robustness and to ensure functionality under all possible variations, it is very important that both the $I_{\mathrm{on},p}/I_{\mathrm{off},n}$ and $I_{\mathrm{on},n}/I_{\mathrm{off},p}$ ratios are high enough. While the latter proved to be no problem in both researched technologies, the first ratio is problematic due to the limited current of the pMOS compared to the nMOS. To increase this ratio, $I_{\mathrm{off},n}$ is reduced through nMOS stacking and $I_{\mathrm{on},p}$ was increased by doubling the width of the pMOS. Table 2 gives the implementation details in both technologies: as can be seen, an additional advantage of TG logic is that (close to) minimal transistor sizing is possible without introducing functionality problems.

Note that in the design of the MAC, all logic gates were implemented differentially to significantly increase the variation-resilience [9]. This comes at only a slight area cost and almost no energy cost because the contribution of the TG logic gates to the total energy consumption is more or less negligible compared to

(a) Half adder. (b) Full adder.

Figure 5: Design of main logic building blocks of the MAC.

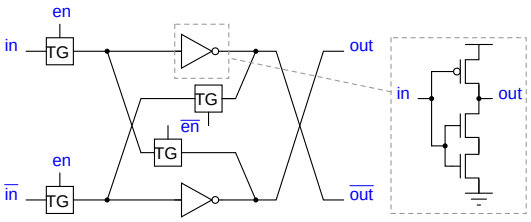


Figure 6: Design of regular latch and inverter.

the one of the latches. By constructing the TG gates differentially, it is also possible to cascade them since they require differential inputs. The higher the number of cascaded logic gates, the more averaging of timing variations is obtained, which is obviously an advantage in the sub-threshold region. However, cascading can also compromise the robustness due to increased signal losses, and the delay is quadratically proportional to the amount of cascading. This trade-off proved to be optimal with 2 cascaded gates [9], therefore the length of a MAC pipeline stage is thus maximally 2 TG logic gates in series.

The main logic building blocks for the MAC are the half and full adders. A half adder (HA) can be easily implemented within the stage length boundaries (see Fig. 5(a)), but for a full adder (FA), this is a challenge. When only using logic gates with 2 inputs, the minimal logic gate depth of a FA is 3. Fortunately, one 3-input logic gate is possible with a single TG logic block: a multiplexer. Therefore, it was possible to satisfy the stage length boundary of 2 with the implementation of the FA (Fig. 5(b)).

5.2. Regular Latch

Fig. 6 shows the schematic of the latch that is used in the pipelined architecture of the MAC. It consists of a fully differential latch, constructed with a minimal amount of inverters, to ensure the regeneration of the signal levels while minimizing energy consumption. The differential nature of the latch adds to the variation-resilience of the total design, which is due to the fact that chances are much lower that variations will compromise the correct interpretation of two complementary inputs than of a single input. The latches are controlled by non-overlapping clock signals (*en_a* and *en_b*) to avoid race problems. The inverter used in the latch is a stacked nMOS inverter [21]. In this case, nMOS stacking is employed to reduce the drive

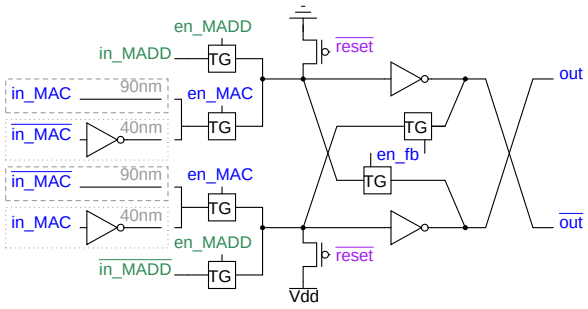


Figure 7: Design of feedback latch.

current $I_{\mathrm{on},n}$, as opposed to the TG logic gates where the decreased leakage was the main reason. Because the pMOS transistor is extremely weak compared to the nMOS in the sub-threshold region for the given technologies, reducing $I_{\mathrm{on},n}$ allowed to relax the relative pMOS sizing to 6 for the 90 nm and 5 for the 40 nm technology.

Table 2 gives the transistor sizing that was used for the inverter in the latch. The sizing was not chosen minimally for several reasons. In both technologies, Monte Carlo simulations showed that in some cases the output signal was not stable even though the latch was locked. This can be explained by unwanted leakage paths through the logic gates which were connected to the output of the latch and by the intra-die variations which severely weakened one of the inverters in the cross-coupled part of the latch. Normally, such cross-coupled inverters are advantageous because they regenerate the input signals and really pull their levels to the supply rails, whereas a single inverter would simply amplify the output signals. However, when one of these cross-coupled inverters becomes too weak due to variations, the feedback in the loop actually accelerates an unwanted bit flip. A solution for this problem is to upsize the inverter, which increases the drive strength of the inverter and reduces its sensitivity to variations. Another reason was that it is imperative that the latch always interprets its input signals correctly, under all possible variations. Upsizing helps again in this case because of the decreased variability. The amount of upsizing was then determined by calculating the probability of failure of a latch under variations: the distribution of the output level of a signal that was propagated through a chain of TGs was compared to the distribution of the offset voltage of the latch:

$$P(\text{failure}) = \int_x P(\text{level}_{\text{out}} = x).P(V_{\text{offset}} > x).\mathrm{d}x \tag{6}$$

In the 90 nm technology, upsizing with a factor 1.5 proved to be sufficient, but due to the increased variability, the 40 nm version needed a slightly higher factor of 2.

5.3. Feedback Latch

As explained in Section 4, the feedback latches, which are placed on the main diagonal of the MAC, enable operation in 3 different modes of the MAC. Fig. 7 shows the schematic of the FB latch. The elements

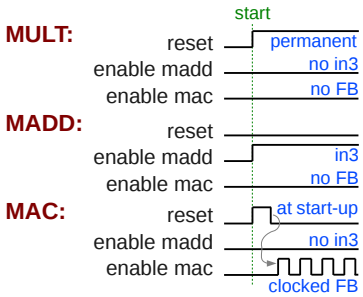


Figure 8: Timing diagram of the different operating modes.

of the FB latch are identical to the ones of the regular latch, except for the reset transistors. The reset is not performed by a single transistor at one side of the cross-coupled inverters because this would lead to ratioed design, which is to be avoided in sub-threshold circuits due to the high sensitivity to variations. Therefore, reset has to be performed at the two sides of the inverters without cross-coupling them through the TGs. A differential reset requires a pull-up and a pull-down transistor on each side respectively (Fig. 7). An important consideration is the leakage contribution of the reset mechanism, since the storage functionality of the latch can be disturbed by this leakage. A minimal pMOS is thus chosen at both sides to reduce leakage. Such a minimal pMOS leaks significantly less than a minimal nMOS transistor in both technologies (e.g. 12.6 times less at $V_{\mathrm{dd}} = 150\,\mathrm{mV}$ in the 90 nm technology).

The timing configuration per operation mode of the FB latch can be seen in Fig. 8:

- MAC mode: The clock signal *en_MAC* is configured to be complementary to *en_fb*. The MAC input bits are the previously calculated product bits out_{prev} used for accumulation and the MADD inputs bits are cut off (*en_MADD*=0). At start-up of the MAC mode, it is compulsory that the FB latches are reset to ensure the first accumulation with zeroes.

- MULT mode: The FB latches are permanently reset while all input bits are cut off (*en_MAC*=0 and *en_MADD*=0), so that addition with 0 is ensured. The reset mechanism is slightly complicated to remove any ratioed design: at start-up, the reset is pulled high, while *en_fb* is kept low. After a certain amount of time when it is sure that the reset node signal levels have settled, *en_fb* is pulled high to establish the regeneration characteristic of the cross-coupled inverters and ensure that signal levels are full-swing.

- MADD mode: Feedback is cut off (*en_MAC*=0) and a third 32-bit input C is provided from the right side (Fig. 3(b)). Due to the pipelining, these MADD input bits need to be delayed by placing latches to ensure arrival to the FB latches on the correct moment. To insert C, the signal *en_MADD* is permanently high, while *en_fb* is always low.

When going to the 40 nm technology node, a change in the topology of the FB latch was necessary:

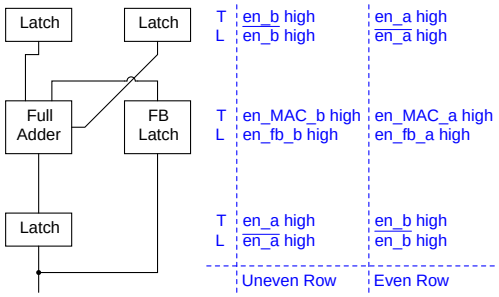


Figure 9: Zoomed in part of the diagonal accumulation of the MAC, with the timing signals of the throughput (T) and locked (L) phases of the latches and the feedback latch added according to the row.

inverters were added at the MAC inputs (Fig. 7). Although various measures were taken to cope with the increased variations in the timing block (which will be addressed in Section 6), in a few rare cases of intra-die simulations a timing error still occurred. Fig. 9 shows the detailed configuration of the FB latch in the diagonal accumulation structure of the MAC. The timing signals change according to the row because of the non-overlapping clock signals *en_a* and *en_b*. The situation where the problem occurred is the following for an uneven row: the latch below has just locked and its output signals are full-swing, as wanted. Then, the FB latch goes transparent and out of lock, but there is a slight 1-1 overlap between *en_MAC_b* and *en_fb_b*. The FB latch is thus very briefly transparent and in lock simultaneously. If the regular (REG) latch is accidentally weaker than the FB latch due to mismatch and the bits saved in both are different, this 1-1 overlap can occur long enough so that kickback takes place and the stored bit in the FB latch interferes with the stored bit in the REG latch, causing it to flip. The most convenient solution to avoid this unwanted kickback is to insert inverters between the outputs of the REG latch and the inputs of the FB latch, hence the kickback will never be able to cause a bit flip in the locked REG latch. This increases the energy consumption of the FB latch, but is necessary to reduce its variation sensitivity and to increase the total yield. Moreover, for an N-bit MAC, only N feedback latches are required. In comparison to the large amount of REG latches, this solution has only a very limited impact on the total energy consumption.

6. Timing

6.1. General Functionality

Fig. 10 shows the implementation of the timing used for the MAC. There are 3 inputs for the timing: the input clock *clock_in* from which the non-overlapping clocks are deduced, and the 2 previously mentioned configuration bits *reset_in* and *madd_in* to determine the operation mode (explained in Table 3). The outputs consist of the clock signals for the REG latches *en_a* and *en_b*, clock signals *en_MAC_a/b* and *en_fb_a/b* for the FB latches, as well as the enable signal for the MADD mode *en_MADD* and the reset signal $\overline{reset}$.

The internal signals *select_fb* and $\overline{reset}_{delayed}$ are used to configure the timing signals of the FB latch.

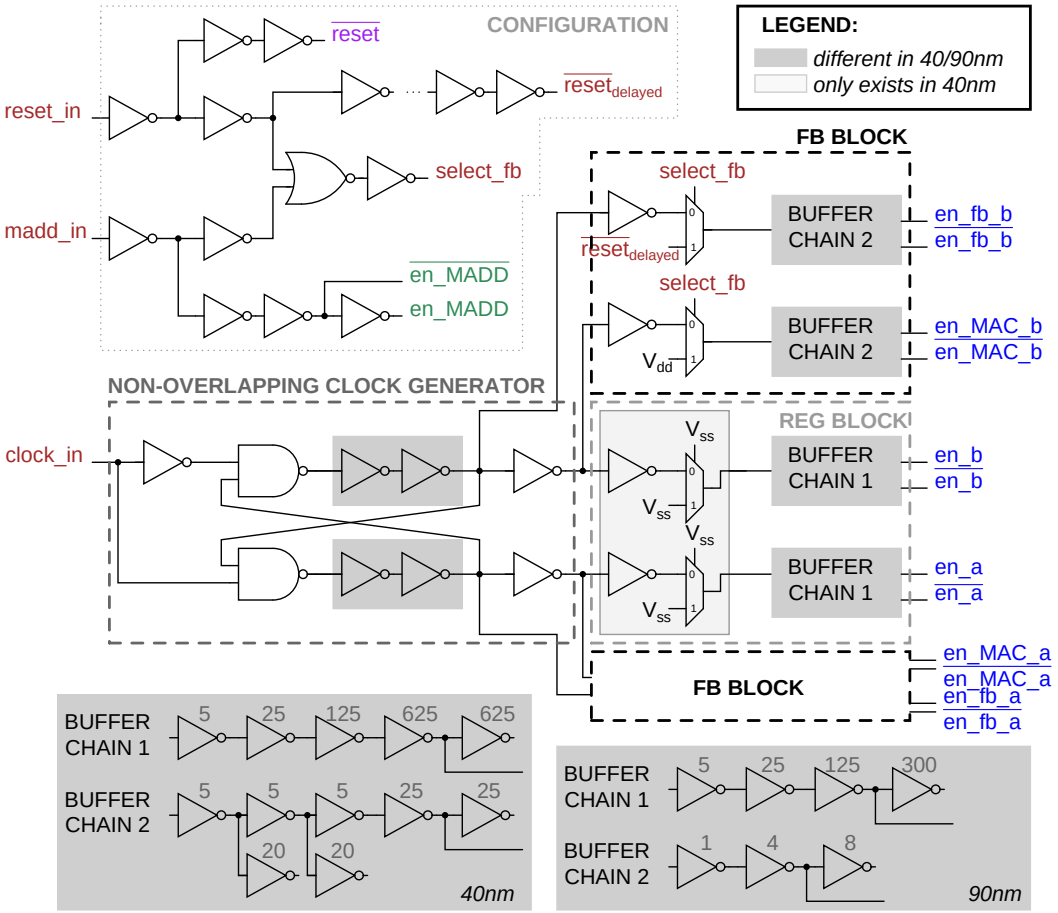


Figure 10: Implementation of the timing.

Operation mode	*reset_in*	*madd_in*
MAC	0	0
MULT	1	0
MADD	0	1

Table 3: Configuration bits per operation mode.

The amount of delay that is inserted for the $\overline{reset}_{delayed}$ is determined by process corner simulations. As explained before, this inserted delay needs to ensure that the signal levels of the reset nodes of the FB latch are settled before establishing the cross-coupled connection. Therefore, the delay of the inverter chain is determined to be higher than the maximal rise and fall time of these nodes in all process corners.

The timing block functions at the same supply voltage V_{dd} as the MAC. The logic gates used in the timing are implemented as standard CMOS logic gates and not as TG logic gates because differential input signals were not available and only a few logic gates were needed. More precisely, the NAND gate is implemented as a regular standard CMOS NAND with appropriate (and therefore increased) sizing of the pMOS, and the NOR gate is implemented using stacked nMOS transistors to reduce the required pMOS sizing.

6.2. Technology Differences

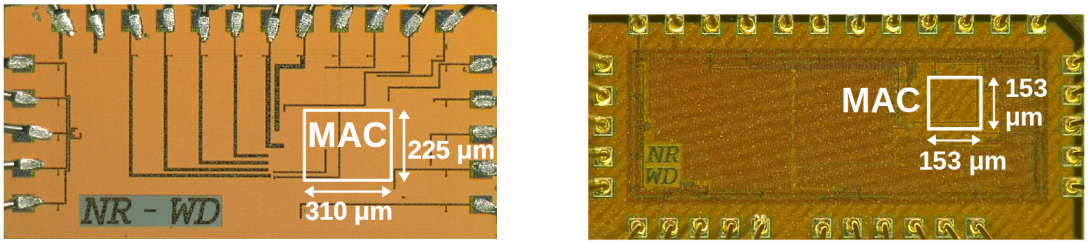
Some of the blocks of the timing are implemented different in both technologies. This comes from the significant increase in variations in the 40 nm technology, which introduced many extra challenges. The 90 nm technology was significantly less sensitive to variations. Note that all inverters of which the sizing is not explicitly mentioned in Fig. 10 were implemented minimally (a relative sizing of 1) in the 90 nm node, whereas in the 40 nm case they were implemented with a relative sizing of 5 to reduce the sensitivity to variations. This upsizing is only used in the timing block, the MAC implementation is sized according to the details provided earlier in Table 2.

The main considerations that had to be taken into account when designing the timing were:

- **Ensure the non-overlap time between the REG clock signals:** The non-overlap time between *en_a* and *en_b* is controlled by the non-overlapping clock generator. The grey shaded area indicates the inverter chain that is inserted to increase the non-overlap time of the clock signals. So that under all variations there would never occur any overlap, a chain of 4 respectively 6 inverters was sufficient for 90 nm and 40 nm.

- **Ensure the non-overlap time between the FB clock signals:** In MAC mode, the FB latches work as REG latches and therefore a non-overlap time between *en_MAC* and *en_fb* is required. This is ensured by matching the paths of both signals as good as possible, which was obtained by having the exact same amount of logic gates and inverters in both paths. To reduce mismatch, the muxes in the FB block are implemented as TG muxes. This was possible because only the controlling signal of a TG mux needs to be differential (Fig. 4).

- **Ensure the non-overlap time between the REG and the FB clock signals:** Since the FB latches work as regular latches in MAC mode, it is imperative that there is no overlap between *en_a/b* and *en_MAC_b/a* (refer also to Fig. 9). In the 90 nm technology, this proved to be not an issue. In the 40 nm node however, matching of the paths was crucial to satisfy this requirement. A few measures were taken to match the paths of the regular and the FB clock signals as meticulous as possible. First of all, dummy inverters and muxes were inserted in the regular path (Reg Block in Fig. 10) to match the delay of the same elements in the FB Block. Additionally, buffer chain (BC) 2 of the FB Block needed to be matched to BC 1 of Reg Block. Simply increasing the sizing of the buffers would unnecessarily increase the energy consumption. Therefore, BC 2 was matched by adding buffers that increased the fan-out of the previous buffer but that were not present in the signal path, thus allowing to not increase the size of the following buffer.



(a) 90 nm MAC. (b) 40 nm MAC.

Figure 11: Chip micrographs.

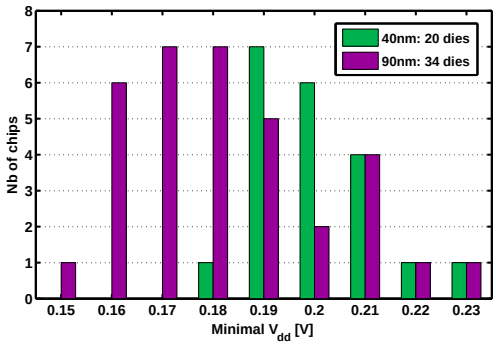


Figure 12: Distribution of the minimal functional supply voltage of the measured dies.

7. Measurement Results

Fig. 11 shows the micrographs of both chips. To acquire an as dense layout as possible for the MAC, the dedicated tool Datapath Generator (DPG) [22] was used, which is very useful for place and route of semi-regular big structures. As inputs, DPG requires a description of the system in structural VHDL, as well as a leaf-cell library. This library contains the custom-made physical layouts of the different gate-level building blocks that are used in the system. Since the timing blocks were too irregular, their layout was carried out full-custom. The resulting active area of the 90 nm version is 225x310 μm^2, while the 40 nm one is 153x153 μm^2. This corresponds to an area reduction of 66 %. According to the classical scaling law, a reduction of around 80 % is expected, but this law does not apply anymore since the transistor area and wiring pitch do not scale likewise. Scaling according to transistor area results in a reduction of 50 % and according to wiring pitch −44 %. To conclude, the 40 nm version has scaled exceptionally well compared to the 90 nm version.

All measurement results provided below are for the MAC mode. Measurements of the two other modes produce very similar results. To study the variation-resilience of both designs, a significant number of dies was measured in both cases: 34 dies for the 90 nm case and 20 for the 40 nm case. All important specifications of the measurement results are compared in Table 4.

Fig. 12 provides the measured minimal supply values of all dies at which both chips were still functional.

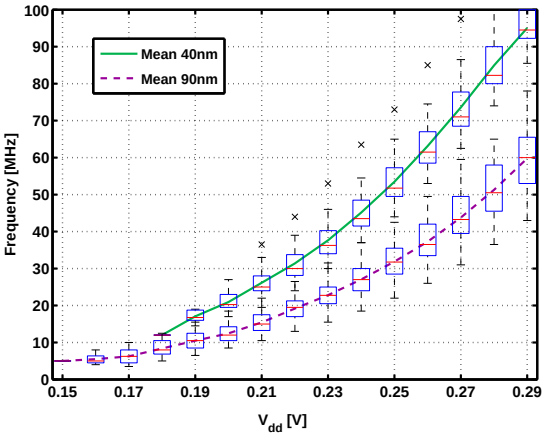


Figure 13: Boxplot of the measured maximum clock frequency as function of V_{dd}.

Measurement results show that $V_{\mathrm{dd},min}$ of the 90 nm MAC is 150 mV, and the 40 nm MAC is able to work down to 180 mV. The measurements thus demonstrate a very good match with the practical $V_{\mathrm{dd},min}$ values of 158 mV and 186 mV respectively, which were obtained in Section 2.

Fig. 13 shows the measured maximum operating frequencies at which each die was able to function correctly at a certain V_{dd}. At 190 mV, a mean clock frequency of 17.06 MHz is obtained with the 40 nm MAC, which is 63 % higher than the mean frequency of 10.48 MHz of the 90 nm MAC at that supply. For a supply ranging from 190 to 290 mV, the 40 nm MAC achieves a mean clock frequency improvement of approximately 66 % over the 90 nm MAC. In the same supply range, the mean variation σ/μ is 11.93 % for 40 nm and 16.77 % for 90 nm, thereby illustrating the variation-resilience of both designs. To conclude, the 40 nm MAC is able to operate significantly faster than the 90 nm MAC for the same supply voltages. Moreover, it also achieves a better variation-resilience than the 90 nm MAC, with a reduction of 29 %. This can be explained by the fact that for the same V_{dd}, the 40 nm transistors are further away from the sensitive sub-threshold region due to their lower V_{T}. Finally, these measurement results thus show that technology scaling is beneficial for sub- and near-threshold circuits in terms of clock frequency. Note also that the expected decrease of delay is accomplished, as opposed to what the simulations suggested (see also Section 3).

The total energy consumption per MAC operation is shown in Fig. 14. Extra on-chip circuitry makes it possible to do at-speed energy consumption measurements while applying arbitrary inputs. The minimum-energy point (MEP) of both designs coincides at 190 mV, where the 40 nm MAC consumes 1.32 pJ per operation, which is an increase of 51 % compared to the 0.87 pJ of the 90 nm MAC. For the 190 to 290 mV supply range, the energy consumption of the 40 nm version is 46 % higher than the 90 nm MAC. Whereas the variation-resilience still improves, unfortunately the energy consumption deteriorates considerably. This total energy consumption consists of a static and a dynamic component. As discussed in section 3, the

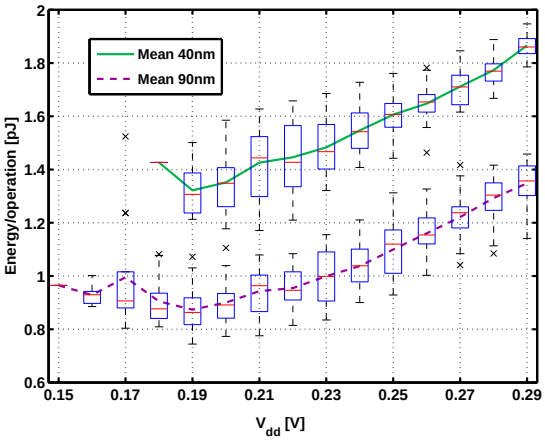


Figure 14: Boxplot of the measured energy consumption per operation as function of V_{dd}.

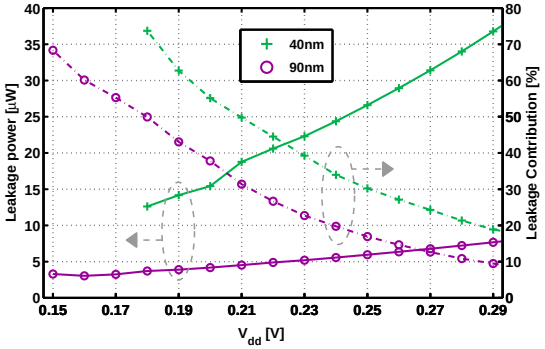


Figure 15: Measured leakage power and contribution of the leakage to the total power consumption as function of V_{dd}.

dynamic energy is expected to decrease with scaling, and the static or leakage energy will increase. In this comparison, the dynamic energy scaling does not completely follow the ideal scaling laws, as the sizing of the basic building blocks (see Table 2) and the timing (see section 6.2) is changed. Moreover, in reality, technologies do not scale according to the ideal scaling laws e.g. wire capacitances do not scale as well as transistor capacitances. Regarding the leakage component, Fig. 15 provides the measured absolute leakage power and relative contribution to the total power consumption as function of V_{dd}. The increased total energy consumption can mainly be attributed to the increased leakage. At a supply of 190 mV, the leakage contribution of the 90 nm MAC to the total power consumption is 43 %, while the 40 nm MAC is dominated by a leakage contribution of 63 % at that point. Moreover, not only the relative leakage contribution increases, but the absolute leakage power increases drastically as well, e.g. from 3.9 µW to 14.2 µW at 190 mV. The leakage component in the total power consumption thus increases substantially for the same supply voltage, thereby explaining the increased total energy consumption.

The question is now: is it advisable to go to advanced nanometer technologies for ultra-low-voltage designs? From an area perspective, it certainly is. Furthermore, the operating frequency increases drastically,

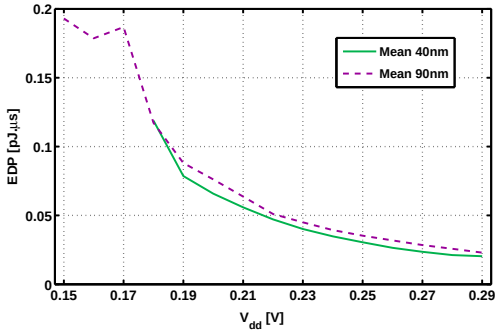


Figure 16: Measured Energy-Delay Product as function of V_{dd}.

but so does the energy consumption per operation. To be able to make a fair comparison between these last two metrics, the Energy-Delay Product (EDP) is the adequate Figure-Of-Merit (Fig. 16). The EDP was calculated as the energy consumption per operation divided by the clock frequency (or throughput) of the MAC. It is not calculated by multiplying energy by the total latency, since the throughput is the metric that indicates the number of inputs that can be calculated in a certain time period for pipelined systems. In terms of EDP, the 40 nm design outperforms the 90 nm version, with a reduction of 13 % for the 190 to 290 mV supply range. To conclude, in an application where energy consumption is of vital importance and speed is of much lower concern, the 90 nm version is the more suitable technology of both, whereas from an area point of view, the 40 nm version is recommended. From an EDP perspective, it depends on whether the ultra-low-voltage design is used in an application or a larger system with a fixed supply voltage, where the 40 nm version performs better at a single given supply voltage, or whether the supply voltage can be freely chosen. In the latter case, it is possible to operate the 90 nm MAC at the same frequency and a slightly higher supply voltage as the 40 nm version, while achieving a lower energy consumption and hence a lower EDP.

Previous work also predicted that the static energy increases dramatically with technology scaling and that this trend can compromise scaling benefits. However, the prediction of the technology node at which this compromising point will occur differs. As stated in [23], the minimum energy consumption increases for the same design when going from a 90 nm technology to a 45 nm node. In [6], it is found that the static energy increase will specifically be dramatic at the 32 nm node and that the benefit of scaling in terms of energy consumption will start to diminish for 45/32 nm technology nodes and below. The authors of [5] stated that by scaling technology from 0.25 µm to 65 nm, the energy consumption can be reduced significantly, but that from an energy consumption point of view, there is no clear benefit to use technologies smaller than 45/65 nm for ultra-low-power purposes. The authors predict that the EDP will start to slowly increase at the 32 nm node.

With our measurement results, we can conclude that, in terms of the energy-performance trade-off, ultra-low-voltage circuits should be scaled down to advanced nanometer technologies, at least until the

CMOS Technology		90 nm	40 nm	**Difference**
Active area	[μm^2]	225x310	153x153	−66 %
Minimal V_{dd}	[mV]	150	180	+20 %
Clock frequency	[MHz]			
@ Minimal V_{dd}		5.0	12.0	
@ 190 mV		10.48	17.06	+63 %
@ 250 mV		31.88	53.48	+68 %
σ/μ @ 190-290 mV		16.77 %	11.93 %	−29 %
Energy/operation	[pJ]			
@ Minimal V_{dd}		0.97	1.43	
@ 190 mV	(MEP)	0.87	1.32	+51 %
@ 250 mV		1.10	1.61	+46 %
σ/μ @ 190-290 mV		7.94 %	6.10 %	−23 %
EDP	[pJ.µs]			
@ 190 mV		0.088	0.079	−11 %
@ 250 mV		0.035	0.031	−14 %
Leakage power	[µW]			
@ 190 mV		3.90	14.18	+364 %
@ 250 mV		5.96	26.60	+447 %

Table 4: Comparison of measurement results.

40 nm node. The benefits of further scaling depend on both the increased leakage, as well as the increased variations. The domination of static leakage of ultra-low-voltage designs in advanced nanometer technologies has consequences for the future of scaling. If, for future technologies below 40 nm, the leakage becomes too high compared to the increase in speed, there will not be any improvement in EDP at a given supply anymore for ultra-low-voltage circuits. More precisely, there will come a point in scaling when the gain of the decreased dynamic energy consumption will be outweighed by the increase in static leakage and the increased variability. If the advance in speed is not able to compensate this, the EDP will not reduce and there will be nothing to gain from further scaling. In this 40 nm technology, the EDP improves because the balance between speed gain, dynamic energy reduction and static leakage increase is still positive.

8. Conclusion

This paper studied the effect of technology scaling on variation-resilient sub-threshold circuits through different approaches. First, a practical expression of the minimum feasible supply voltage for a digital circuit was derived. As such, a theoretical minimum supply of 101 mV was calculated. For a certain CMOS technology, this equation makes it possible to calculate a practical limit for $V_{dd,min}$. Second, a test case of an extensive digital circuit, i.e. a Multiply-Accumulate unit, was designed, processed and measured in a 90 nm and a 40 nm CMOS technology. The measurement results were extensively compared in order to study the benefits of technology scaling for ultra-low-voltage circuits. The measurements demonstrate a drastically improved operating frequency at the cost of a higher energy consumption, resulting in a reduced

Energy-Delay Product at a given supply voltage for the 40 nm MAC. Scaling to the 40 nm node is beneficial for ultra-low-voltage designs in terms of area and of EDP at a fixed V_{dd}, but not in terms of energy consumption. It is shown that the effect of scaling on the EDP for such designs will be positive as long as the static leakage is kept under control. Although advanced nanometer technologies suffer from an increased variability, measurements show that both MACs are still variation-resilient.

Acknowledgment

This work is supported by the Research Foundation - Flanders (FWO). The authors would like to thank Tobias Noll, Oliver Weiss and Michael Meixner from RWTH Aachen University for the opportunity to work with Datapath Generator and the provided support.

References

[1] M. Alioto, Ultra-low power VLSI circuit design demystified and explained: A tutorial, IEEE Trans. Circuits and Systems I 59 (1) (2012) 3–29. doi:10.1109/TCSI.2011.2177004.

[2] S. Hanson, M. Seok, D. Sylvester, D. Blaauw, Nanometer device scaling in subthreshold logic and SRAM, IEEE Trans. Electron Devices 55 (1) (2008) 175–185. doi:10.1109/TED.2007.911033.

[3] B. Paul, A. Raychowdhury, K. Roy, Device optimization for digital subthreshold logic operation, IEEE Trans. Electron Devices 52 (2) (2005) 237–247. doi:10.1109/TED.2004.842538.

[4] B. Calhoun, S. Khanna, R. Mann, J. Wang, Sub-threshold circuit design with shrinking CMOS devices, in: ISCAS, 2009, pp. 2541–2544. doi:10.1109/ISCAS.2009.5118319.

[5] A. Tajalli, Y. Leblebici, Design trade-offs in ultra-low-power digital nanoscale CMOS, IEEE Trans. Circuits and Systems I 58 (9) (2011) 2189–2200. doi:10.1109/TCSI.2011.2112595.

[6] D. Bol, R. Ambroise, D. Flandre, J. Legat, Interests and limitations of technology scaling for subthreshold logic, IEEE Trans. VLSI Systems 17 (10) (2009) 1508–1519. doi:10.1109/TVLSI.2008.2005413.

[7] M.-E. Hwang, Supply-voltage scaling close to the fundamental limit under process variations in nanometer technologies, IEEE Trans. Electron Devices 58 (8) (2011) 2808–2813. doi:10.1109/TED.2011.2151257.

[8] W. Zhao, Y. Cao, New generation of predictive technology model for sub-45nm design exploration, in: ISQED, 2006, pp. 585–590. doi:10.1109/ISQED.2006.91.

[9] N. Reynders, W. Dehaene, Variation-resilient sub-threshold circuit solutions for ultra-low-power digital signal processors with 10MHz clock frequency, in: ESSCIRC, 2012, pp. 474–477. doi:10.1109/ESSCIRC.2012.6341358.

[10] N. Reynders, W. Dehaene, A 210mV 5MHz variation-resilient near-threshold JPEG encoder in 40nm CMOS, in: ISSCC, 2014, pp. 456–457. doi:10.1109/ISSCC.2014.6757511.

[11] R. Swanson, J. Meindl, Ion-implanted complementary MOS transistors in low-voltage circuits, IEEE J. Solid-State Circuits 7 (2) (1972) 146–153. doi:10.1109/JSSC.1972.1050260.

[12] A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, E. Nowak, Low-power CMOS at Vdd = 4kT/q, in: Device Research Conference, 2001, pp. 22–23. doi:10.1109/DRC.2001.937856.

[13] A. Wang, B. Calhoun, A. Chandrakasan, Sub-threshold Design for Ultra Low-Power Systems, Springer, 2006.

[14] N. Weste, D. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, 4th Edition, Addison-Wesley, 2011.

[15] Y. Tsividis, C. McAndrew, Operation and Modeling of the MOS Transistor, 3rd Edition, Oxford University Press, 2011.

[16] J.-J. Kim, K. Roy, Double gate-MOSFET subthreshold circuit for ultralow power applications, IEEE Trans. Electron Devices 51 (9) (2004) 1468–1474. doi:10.1109/TED.2004.833965.

[17] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits, Proceedings of the IEEE 91 (2) (2003) 305–327. doi:10.1109/JPROC.2002.808156.

[18] M. Pelgrom, A. C. J. Duinmaijer, A. Welbers, Matching properties of MOS transistors, IEEE J. Solid-State Circuits 24 (5) (1989) 1433–1439. doi:10.1109/JSSC.1989.572629.

[19] N. Reynders, W. Dehaene, A 190mV supply, 10MHz, 90nm CMOS, pipelined sub-threshold adder using variation-resilient circuit techniques, in: A-SSCC, 2011, pp. 113–116. doi:10.1109/ASSCC.2011.6123617.

[20] M. Hatamian, G. Cash, A 70-MHz 8-bitx8-bit parallel pipelined multiplier in 2.5-μm CMOS, IEEE J. Solid-State Circuits 21 (4) (1986) 505–513. doi:10.1109/JSSC.1986.1052564.

[21] N. Reynders, W. Dehaene, Variation-resilient building blocks for ultra-low-energy sub-threshold design, IEEE Trans. Circuits and Systems II 59 (12) (2012) 898–902. doi:10.1109/TCSII.2012.2231022.

[22] O. Weiss, M. Gansen, T. Noll, A flexible datapath generator for physical oriented design, in: ESSCIRC, 2001, pp. 393–396.

[23] D. Bol, D. Kamel, D. Flandre, J.-D. Legat, Nanometer MOSFET effects on the minimum-energy point of 45nm subthreshold logic, in: ISLPED, 2009, pp. 3–8. doi:10.1145/1594233.1594237.